DOCUMENT RESUME

ED 334 234 TM 016 751

AUTHOR	Fitzgerald, Nicholas B.
TITLE	Program Improvement through Evaluation: Connecticut's First Round of Sustained Effects Studies. Feedback, Occasional Paper Number 1.
INSTITUTION	RMC Research Corp., Hampton, N.H.
PUB DATE	Aug 86
NOTE	19p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Compensatory Education; Data Analysis; Educational Research; Elementary Secondary Education; *Evaluation Methods; Program Development; Program Evaluation; *Program Improvement; *Quality Control; *Research Design; Research Problems; Test Interpretation; *Test Use
IDENTIFIERS	*Connecticut; Education Consolidation Improvement Act Chapter 1; Quality Indicators; *Sustaining Effects Study

ABSTRACT

A quality control review of Connecticut's Sustained Effects studies was undertaken by the Chapter 1 Evaluation Technical Assistance Center (TAC) operated by RMC Research Corporation. Nineteen Connecticut school districts had selected from among four basic evaluation designs in implementing the Chapter 1 Sustained Effects requirement during 1982 to 1985. These designs enabled districts to examine the long-term effects of their Chapter 1 programs on: (1) summer drop-off students; (2) continuing students; (3) exiting students; and (4) continuing versus exiting students. The review identified the following minor areas for improvement in future Sustained Effects studies: evaluation design; test administration; and analysis of appropriate test scores. Most of the studies were of adequate technical quality and appeared to be useful. Areas of major significance in which improvement was needed were: (1) framing evaluation questions; (2) data analysis and interpretation; and (3) development of action plans for program implementation. Solutions to these problems are recommended. (SLD)

FEEDBACK

Occasional Paper Series

Program Improvement Through Evaluation:

Connecticut's First Round of Sustained Effects Studies

(Paper Number 1)

Nicholas B. Fitzgerald, Ph.D.
Chapter 1 Evaluation Technical Assistance Center
RMC Research Corporation
Hampton, NH

August 1986

Executive Summary

Connecticut school districts selected from among four basic evaluation designs in implementing the Chapter 1 Sustained Effects requirement during the period 1982-1985. These evaluation designs enabled local districts to examine the long term effects of their Chapter 1 programs.

A review of these Sustained Effects studies by the Chapter 1 Evaluation Technical Assistance Center was undertaken as a quality control measure, the result of which was the identification of areas where future Sustained Effects evaluations might be improved. Minor areas identified for improvement included the following:

- o Evaluation Design
- o Test Administration
- o Analysis of Appropriate Test Scores

Areas of major significance in which Sustained Effects evaluations could be improved centered on issues of evaluation utilization and associated technical aspects of data quality. These areas include:

- o Framing Evaluation Questions
- o Data Analysis and Interpretation
- o Development of Action Plans for Program Improvement.

Solutions to the identified problems are recommended in the concluding section of this paper.

INTRODUCTION

A quality control review of Connecticut's Sustained Effects studies was undertaken by the Chapter 1 Evaluation Technical Assistance Center (TAC) operated by RMC Research Corporation. The purpose of this paper is to summarize the findings of that review so that future Sustained Effects evaluations might be improved. Toward that end, the information contained in this report served as a basis for designing the Sustained Effects Workshops conducted in North Haven and Hartford on March 24 and 27, 1986. The findings of this report are also being considered by state administrators in their development of policy governing the Sustained Effects requirement which is anticipated to be disseminated this fall.

The quality control analysis examined eight areas of concern in the design, implementation, and reporting of Sustained Effects evaluations. These areas of concern are summarized by design type in Table 1 (see back page).

Difficulties in planning a Sustained Effects study varied by design type. The easiest type of study to conceptualize appeared to be Design 2: the long-term effect of Chapter 1 on continuing students. Design 1 (Summer Drop-Off), Design 3 (Exiting Students) and Design 4 (Continuing vs. Exiting Students) appeared to be more difficult to plan. These latter Sustained Effects studies encountered design problems related primarily to the use of invalid testing cycles.

Major problems common to all design types included: (a) small sample sizes, (b) a project's capability to increase sample size adequately, and (c) the ability to correctly analyze and interpret the results of a Sustained Effects study. Less problematic, but still of major concern, was the ability to generate a clear and appropriate evaluation question. Finally, the quality control review raises some concern about the local

utilization of evaluation information. To some extent, there was a tendency to ignore negative results. However, in a larger sense, Sustained Effects studies would probably be more useful at the local level if more projects were to incorporate program evaluation into the planning process.

The findings of TAC's quality control review of 119 Sustained Effects studies conducted during the 1982-1985 period are discussed in greater detail in the following two sections. These two sections are descriptive for the most part, the first dealing with areas of minor concern, and the second focusing on areas of major concern. The final section of the paper is more prescriptive in nature and offers some recommendations for improving future Sustained Effects evaluations in both major and minor areas of concern.

AREAS OF MINOR CONCERN

In this section, three quality control indicators are discussed in terms of the nature and extent to which Sustained Effects studies used valid research designs, appropriate testing cycles, and appropriate test scores. The vast majority of Sustained Effects studies did not encounter problems in these areas, but enough districts did have problems in this area of quality control to warrant discussion here.

Overall, for example, 89% of the Sustained Effects studies used valid research designs in which testing cycles were appropriate to those designs, and in 93% of the studies appropriate test scores were used in the analysis of Sustained Effects data. By contrast, the percentages are somewhat lower for those quality control indicators discussed in the section on areas of major concern. We turn now to a discussion of those quality control indicators where only a small minority of Sustained Effects studies experienced difficulty.

Research Design

Design 3 (Exiting Students) and Design 4 (Exiting vs. Continuing Students) appear to be the most difficult to plan, judging by the occurrence of flaws in research design. Typically, the quality control problem in research design involves the use of an inappropriate testing cycle in which the testing plan does not match the evaluation question or the intended research design.

Other manifestations of design flaws included the following:

- o Using different testing cycles for different grades
- o Inability to deliver the proposed design
- o Invalid group comparisons

The last two points require further elaboration. The inability to deliver a proposed design is reflective of the degree of control which many local school districts have (or, more accurately, do not have) over the Chapter 1 population and from whom they intend to collect data. Student attrition and inability to predict student exit or continuation patterns are common reasons for failing to deliver a proposed design. However, in most cases, these studies could have been re-designed during the sustained effect year in order to examine issues related to the program's current situation rather than choosing to abandon the study. This situation has implications at the state level for the provision of technical assistance in cases where a sustained effects study needs to be redesigned.

The problem of invalid group comparisons was specific to Design 4 (Continuing vs. Exiting Students), particularly when the evaluation question dealt with length of treatment issues (e.g., one vs. two years participation in Chapter 1). In this design context, both groups need to exit the program before a valid comparison can be made in addressing policy concerns related to length of stay in the program. In essence, Design 4 does not

accommodate evaluation questions bearing on policy issues related to the differential effects of length of participation in Chapter 1.

Testing Plan

In about ten percent of the Sustained Effects studies, either an inappropriate testing schedule was used (which invalidated the evaluation design), or tests were administered outside of the correct norming dates. The more common type of norming date violation was one in which the Sustained Effects test was given too early and which would tend to underestimate the impact of a program. Use of incorrect test levels was also involved, but the extent and nature of this problem is masked by the lack of information on this topic in the reports submitted.

Test Score

The use of test scores inappropriate for Chapter 1 evaluation occurred in a few of the Sustained Effects studies. While not a major problem, the use of appropriate test scores is a fundamental consideration to the norm-referenced evaluation model. The most common manifestation of this problem was the use of raw scores, percentile ranks, or grade equivalents in the analysis of Sustained Effects data rather than an equal interval scale metric like the normal curve equivalent (NCE) score.

There were also a few cases in which noncomparable tests and test scores were used across the three testing points. For example, an IQ test might be used during the pretest-posttest period and then an achievement test would be employed at the Sustained Effects data point. This situation would result in using changes in IQ scores to describe the effect of the program and then using a different, noncomparable test score (e.g., an achievement test scores) to draw conclusions about sustained effects. This

situation is analogous to using different tests and test scores in the same evaluation when those tests and tests scores have not been equated.

AREAS OF MAJOR CONCERN

The majority of Sustained Effects studies exhibited adequate technical quality and appeared to be useful at the local level. Nevertheless, certain types of problems were encountered by a number of projects in conducting a Sustained Effects evaluation. The biggest problem, by far, involved the analysis of Sustained Effects data and related issues of sample size. Another area of major concern is the framing of clear and appropriate evaluation questions. Finally, there is some concern over the utilization of Sustained Effects evaluation information and the degree to which study recommendations are responsive to the results of a Sustained Effects inquiry. We turn now to a more detailed discussion of these topics.

Data Analysis

Local district personnel appear to encounter major difficulties in the analysis of Sustained Effects data in terms of both establishing the nature of a program effect for the base year as well as determining whether the effect is maintained during the Sustained Effects period. This problem is compounded by lack of consensus on the use of standards and decision rules for interpreting Sustained Effects data. In addition, local evaluators tend to limit their analyses to a description of the data rather than to draw conclusions regarding the sample group from which inferences to the Chapter 1 population need to be made.

There was also a tendency to interpret results for individual students rather than examining group data -- the latter rather than the former having implications for the overall program. While the results for individual

students are useful for diagnostic purposes, the primary intent of a Sustained Effects evaluation is to examine how the program is working more generally for the Chapter 1 population under a given set of educational practices and policies.

The interpretation of Sustained Effects data was perhaps most difficult for those studies hampered by small sample size. The problem of adequate sample size affected almost two-thirds of all Sustained Effects studies conducted in Connecticut between 1982-1985. This problem largely concerns the question of "when is a difference a real difference?"

Small Sample Size. All measurement contains some amount of error which influences the degree of consistency or reliability of test scores. The reliability of evaluation results based on test scores is, in turn, influenced by the number of scores used to estimate, for example, a program's effectiveness. In short, one's confidence in the reliability of conclusions about a program's effectiveness, based on test scores, will increase with sample size. The converse proposition is also true: the smaller the sample size, the less confidence one can place in the reliability of conclusions drawn from test score data.

By the term "small sample size" is meant the use of samples generally less than 25 students upon which a sustained effects analysis was based. The sample size problem tends to be manifested at the grade level and could be compensated for in about 60% of the cases by conducting a pooled analysis (i.e., collapsing across grade levels, as appropriate) which would increase the sample size enough to offset the associated problem of measurement error. To the extent that sample size can be adequately increased, the measurement error problem will be minimized. Otherwise, the solution strategy is to replicate the study so that conclusions may be drawn from a pattern of results.

Replicating a study means to repeat the study. If a small Chapter 1 program conducts the same study twice over a three-year period, and if the pattern of results is the same each time, we can have greater faith in those results than if we relied only on a one-shot approach to program evaluation. If the results of the two studies fail to coincide, we should then exercise caution in drawing conclusions from either study alone. Under such circumstances, further refinements might be warranted in the evaluation plan.

Solutions to analysis and interpretation problems were discussed in the March workshops in terms of options for standard setting and procedures useful to establishing both base-year program effects, as well as the maintenance of program effects. These strategies, and more, are reviewed in the conclusions/recommendations section of this paper.

Framing Evaluation Questions

In the majority of Sustained Effects studies, evaluation questions were framed in a clear and appropriate manner. Adequate evaluation questions were apparently easiest to write for the Summer Drop-Off design and for studies intending to examine the effect of Chapter 1 on students exiting the program.

The presence of unclear and inappropriate evaluation questions occurred most frequently in those study designs focusing on the long-term effects of Chapter 1 on continuing students and in comparisons of exiting versus continuing students. In these two designs in particular, the evaluation questions rarely addressed issues in local compensatory education policy or practice. In these cases, the reader is left wondering how Sustained Effects evaluation results would be used to inform program managers about the worth of current education practices and policies at the local level.

Aside from the issue of policy relevance, approximately one-quarter of the Sustained Effects evaluation questions appeared to be unclear in defining the intent of the study, or the evaluation question was framed in a way that made it inappropriate for a Sustained Effects study. Evaluation questions inappropriate to a Sustained Effects design tended to be those which made no reference to concerns for sustaining program gains subsequent to the base year. For example, generic questions of an exploratory nature (e.g. "What is the rate of growth" for some usually undefined group of students) might be posed in which the time period of interest was often the base year, rather than a subsequent time-frame. Unclear evaluation questions tended to be those which were vague in defining the groups of student involved, specifying grade levels of interest, the time-frames defining the base-year and sustained effects period, and the kind of data which would be collected. In short, unclear evaluation questions leave the reader with only a vague sense of what a Sustained Effects study is trying to accomplish. But perhaps more importantly, vagueness in the framing of the evaluation question is a reflection on the conceptual clarity of the overall evaluation plan, which, in turn, increases the potential risk of conducting a flawed evaluation or a study whose results are of little use to anyone.

The framing of an evaluation question determines, to a large extent, the direction in which a Sustained Effects study should go, which reminds me of Alice's first encounter with the Cheshire Cat in Wonderland. Alice began by asking the Cheshire Cat:

> "Would you tell me, please, which way I ought to walk from here?"
>
> "That depends a good deal on where you want to get to," said the Cat.
>
> "I don't much care where," said Alice.
>
> "Then it doesn't matter which way you walk," said the Cat.

The evaluation question reflects "where you want to get to." But unlike Alice, for those of you who "much care where," the process of framing a good evaluation question will help define what you want to accomplish and how to get there.

Utilization of Evaluation Results

Overall, approximately 80% of the Sustained Effects studies yielded reasonable action plan statements in the final section of the reports reviewed. These statements were written in response to Item 19 on the reporting form which asks, "What changes, if any, will be made in your Chapter 1 program as a result of this study?"

Acceptable action plans can be described as falling into one of three categories in which the evaluation results were interpreted correctly to indicate sustained program effects, nonsustained program effects, or non-significant effects due to small sample size which precluded the drawing of conclusions about program effects.

Sustained Effects. Studies which correctly concluded that the program had produced sustained gains typically recommended no further change to program operations. In rare cases, the evaluation findings were (a) disseminated to LEA staff for planning discussions, or (b) used to reinforce a particular policy if specific program variables had been examined as part of the Sustained Effects study.

Nonsustained Effects. Studies which correctly concluded that the program had not produced sustained effects typically drafted recommendations to do something about the situation. In Summer Drop-Off designs, for example, it was indicated that some form of supplemental reading program would be made available to parents of Chapter 1 students or that the fall semester curriculum would be re-designed to focus on identified problem

areas in order to reinforce skills that students were experiencing difficulty with. In some cases, changes in program evaluation practices were planned, such as the need to institute functional level (i.e., out-of-level) testing.

Nonsignificant Effects due to Sample Size. Studies which correctly concluded that no substantive conclusion was warranted on the basis of small sample size typically recommended that no program changes occur as a result of Sustained Effects evaluation findings. While these LEA evaluations deserve credit for exercising good judgment in recognizing the need for a conservative appraisal of their data, recommendations could have been developed, in some cases, for improving future Sustained Effects studies in ways to compensate for sample size problems. Options for resolving sample size problems are discussed in the concluding section of this paper.

Improving Utilization of Sustained Effects Results

Where action plans could be improved dramatically lies largely in the areas of interpreting Sustained Effects data correctly, drawing conclusions from the data, and in the framing of more useful evaluation questions so that the results of Sustained Effects studies could be brought to bear more directly on policy issues relevant to program improvement. Evaluation data can provide useful and helpful information to program managers only if the data are interpreted correctly, if conclusions can be drawn from the data, and if the data are relevant to issues which are important to local program operations. These topics are discussed below.

Drawing Inferences from Evaluation Data. The sample size problem in many studies created major difficulties for LEA evaluators in the interpretation of their Sustained Effects data. Because the data were analyzed at multiple grade levels in many cases, small sample sizes were created in

which the ability to draw reliable conclusions was not easy, particularly when multiple and often conflicting patterns could be seen in the data. This situation made it difficult to draw conclusions, other than to describe various patterns, and this probably increased the chances for drawing incorrect conclusions -- either false-negative[1], or more typically, false-positive[2] interpretations. The consequences for evaluation utilization of having to wade through such a quagmire of data was that in about 20% of the Sustained Effects studies, incorrect conclusions were drawn in which the recommendations were unresponsive to the real story behind the data.

Sustained Effects studies resulting in false-positive conclusions represent a major concern for the appropriate utilization of evaluation results. In these cases, ineffective programs are promoted as successful Chapter 1 projects in which the typical recommendation is to maintain the status quo. Not infrequently, false-positive reports simply ignore negative results or rationalize negative program effects into positive findings.

When negative findings were apparent, both graphically and statistically inferred, "false-positive" projects were disinclined to accept the results or to utilize such information for planning purposes. Rather, these projects tended to dismiss negative findings, often by blaming the

---

[1] A false-negative interpretation is an incorrect conclusion which asserts that the program failed to sustain achievement gains when in fact program gains were sustained.

[2] A false-positive interpretation is an incorrect conclusion which asserts that the program did sustain base year gains when in fact this was not the case.

test instrument. In short, there was some tendency to resist examining local education practices in light of evaluation data which suggested that programmatic changes might be warranted.

Policy Relevance. The majority of evaluation questions lacked local policy relevance, and at the same time, were framed in such a way that a program's very existence could be threatened if negative results were to be found. In speculating on this state of affairs, it appears that few local projects viewed Sustained Effects evaluation as having much utility or relevance to them other than compliance with state guidelines. This could partially account for a lack of local policy relevance in many of the Sustained Effects evaluation questions, which in turn would tend to inhibit local utilization of evaluation results.

SUMMARY AND RECOMMENDATIONS

The general idea behind the Sustained Effects evaluation requirement is to determine whether Chapter 1 program gains are maintained in the long run. More specifically, Sustained Effects evaluations can be used to determine if program graduates maintain their achievement levels once supplemental compensatory education support has been removed and/or to examine the long-term effectiveness of particular Chapter 1 policies and practices. The Connecticut State Department of Education provides four basic evaluation designs for local districts to select from in order to achieve the above stated purposes of the Sustained Effects evaluation requirement.

The minor problem types discovered in Connecticut's first round of Sustained Effects evaluations (1982-1985) centered around design flaws which could invalidate such a study. These problems were primarily related to test administration issues and the testing design employed. The follow-

ing recommendations summarize these minor problems and at the same time offer some solutions:

- Use comparable or equated tests and test scores, across the base year and sustained effects period.
- Design a testing plan which covers the base year (pretest - posttest) and the sustained effects period (the third data point or sustained effects test).
- Match the testing design to the Sustained Effects evaluation question. Valid testing cycles appropriate for each of the four Sustained Effects evaluation designs include the following:

  Design #1 (Summer Drop-Off)
  Fall - Spring - Fall
  Spring - Spring - Fall

  Designs #2, #3, and #4
  (Exiting, Continuing, or Exiting vs. Continuing Students)
  Fall - Spring - Spring
  Spring - Spring - Spring
  Fall - Fall - Spring

- Administer tests within the empirical norming dates or use interpolated norms if appropriate.
- Administer the test level recommended by the test publisher or conduct functional (out-of-level) testing if appropriate.
- Analyze Sustained Effects data using the Normal Curve Equivalent (NCE) score.

The major problem types found in the first-round Sustained Effects studies generally involved their usefulness to local districts -- the extent to which the results were used in program planning and the extent to which the results were useable in terms of technical quality. With respect to evaluation utilization, results favorable to a program tended to be used to confirm the status quo whereas unfavorable results were often ignored or rationalized away. Only in a minority of cases did programs indicate that they planned to examine local operations on the basis of negative evaluation findings in order to determine how Cnapter 1 services could be made more effective.

The usability of many Sustained Effects evaluations appeared to be influenced not only by the usefulness and relevance of the evaluation questions, but also by the technical quality of the data and the ability to interpret evaluation results correctly. Certainly, future Sustained Effects studies could be made more useful by framing evaluation questions which are more relevant to pertinent Chapter 1 policy issues and which are technically feasible to address at the local level. But even studies with "good" evaluation questions (i.e., relevant to local policy and practice, for which evaluation information can influence action) are not useable if the technical quality of the data is low and/or if the data can not be interpreted correctly.

The main reason for low technical quality of Sustained Effects evaluation data was that sample sizes were too small and thus the data could not yield reliable conclusions. In some cases, the inability to draw conclusions because of small sample size was acknowledged by local districts -- and rightfully so if the sample size problem could not be resolved methodologically. But in over half of the cases where sample size was a problem, it could have been resolved favorably. The following recommendations are offered as solution strategies for those districts encountering problems with sample size:

- Reduce test score attrition thereby increasing the number of students with all three data points
- Increase sample size by conducting a pooled analysis -- collapse across grade levels where appropriate and meaningful to the evaluation question
- Use tests of statistical significance, or employ the Give-or-Take Table to estimate the magnitude of measurement error for a particular sample size:
- Replicate the study when sample size can not be increased -- greater confidence can be placed in two small studies showing similar results

Another means by which Sustained Effects evaluations could be improved is by offering some suggestions on how to interpret the data. The strategy recommended here is to examine group data, first for the base year and then for the sustained effects period. Conclusions need to be drawn, first in regard to Chapter 1 program effects during the base year, and second, determining whether these effects were sustained in the subsequent time period. However, we need a set of standards and some decision rules in order to determine the existence and nature of base year effects, and whether those effects were sustained.

Two options are available for setting base year standards: use of statistical significance testing or appeal to a set standard. Within each of these two approaches to standard setting, two alternatives are available. Districts choosing the statistical significance approach can either conduct a planned comparison test (e.g. the T-test) or use the "give-or-take table" (handed out at the March workshops and also available through the Chapter 1 Evaluation Technical Assistance Center). Districts choosing to appeal to a set standard can use either the current state policy for goal achievement or establish a local performance improvement goal.

State policy for base year standards allows for different goal levels depending on the testing cycle used and the grade levels involved. For districts using a Fall-Spring testing cycle, the state standard for goal achievement is an average gain of 5 NCEs; this standard applies to all grade levels. The standard for districts using an annual testing cycle (Fall-Fall for Spring-Spring) is a gain of 3 NCEs or more for grade levels 2-8 and a gain of at least one NCE for grade levels 9-12.

The decision rules to use in determining the existence of a program effect for the base year depends, of course, upon the standards selected. A program effect will be considered as having been established if one of

the following conditions are met:

- The NCE difference between the mean pretest and mean posttest score is statistically significant.
- The NCE difference between the mean pretest and mean posttest score is large enough to meet the applicable state standard.
- The NCE difference between the mean pretest and mean posttest score shows an improvement over prior local performance.

Once the nature of a program effect has been established for the base year, the final task is to determine whether the base year effect was sustained. The decision rule here is simple and straight forward:

- If the mean NCE difference between the Sustained Effects test and the posttest is greater than or equal to zero, then the program has sustained its base year effects.
- If the mean NCE difference between the Sustained Effects test and the posttest is less than zero, then the base year effect has not been sustained.

Placing a value on the type of base year effect found (e.g., positive, negative, or no effect) and whether sustaining that type of effect is good or bad, is a matter for local judgment and common sense.

Free technical assistance with Sustained Effects issues and problems can be obtained through the Region I TAC by calling 800-258-0802. Consultations, workshops, and written materials on Sustained Effects evaluation and other Chapter 1 evaluation topics can also be obtained at no cost.

In closing, local districts are to be commended for their efforts in this first round of Sustained Effects studies. Hopefully, the feedback provided here will not only improve future Sustained Effects evaluations but will also make them easier to conduct and more useful to local educators.